IDIAP RESEARCH REPORT


# AUTOMATIC TEMPORAL ALIGNMENT OF AV DATA

Danil Korchagin Philip N. Garner John Dines

Idiap-RR-39-2009

DECEMBER 2009

**Centre du Parc**, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

# Automatic Temporal Alignment of AV Data

Danil Korchagin, Philip N. Garner and John Dines
Idiap Research Institute
CH-1920 Martigny, Switzerland
+41 27 721 77 11
{Danil.Korchagin, Phil.Garner, John.Dines}@idiap.ch

## ABSTRACT
In this paper, we describe the automatic audio-based temporal alignment of audio-visual data, recorded by different cameras, camcorders or mobile phones during social events like high school concerts. All recorded data is temporally aligned with a common master track, recorded by a reference camera. The core of the algorithm is based on perceptual time-frequency analysis with a precision of 10 ms. The results show correct alignment in 99% of cases for a real life dataset.

## Categories and Subject Descriptors
F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Non-numerical Algorithms and Problems – *pattern matching*.

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *indexing methods*.

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing – *signal analysis, synthesis, and processing*.

I.1.2 [**Symbolic and Algebraic Manipulation**]: Algorithms – *analysis of algorithms*.

## General Terms
Algorithms.

## Keywords
Audio processing, time-frequency analysis, temporal alignment.

## 1. INTRODUCTION
The TA2 project (Together Anywhere, Together Anytime) is concerned with investigation of how multimedia devices can be introduced into a family scenario to break down technology and distance barriers. In this sense, we are interested in the use of consumer level multimedia devices in novel application scenarios.

One generic scenario is the use of multiple capture devices in a single room. The present investigation concerns the possibility of using multiple consumer level video cameras to record a scene. In a professional scenario, one might expect to be able to use multiple capture devices, and for them all to be synchronised via a common clock or similar [1]. Consumer level devices, however, do not normally provide such capabilities. Further, if the devices are hand-held, we cannot rely in any predictable sense on the video signal. This leaves us with the audio signal from which to infer synchronisation information [2].

In this study, we were provided with a single reference signal from a fixed camera that recorded the whole scene. We were also provided with several auxiliary signals from hand-held cameras that recorded parts of the scene. If we could show that the auxiliary signals could be aligned with the reference signal reliably, then the project could profit from using audio-based temporal alignment. If it were too error-prone or computationally onerous, then other solutions would have to be sought.

## 2. PRE-PROCESSING STAGE
Consider a simple high school concert event. The duration of the corresponding master track can easily be of the order of a small number of hours. This in turn corresponds to a large quantity of raw audio data (stereo at 48 kHz). It is normal is such situations to decrease the search space, retaining only useful information [3] for temporal alignment. Accordingly, we assume from the outset that raw PCM audio data is both too voluminous and too noisy to produce good audio alignment. We suppose that good results might be obtained by lower resolution features such as frame energy. Certainly, a resolution approaching video frame rate is sufficient for the purposes of our application.

Given that our broader application is expected to include Automatic Speech Recognition (ASR), the pre-processing takes the form of a standard feature extraction chain used in ASR. In our work we use Mel Frequency Cepstral Coefficients (MFCC) [4] with a 10 ms frame rate. MFCC is a perceptually motivated spectrum representation that is widely used not only in speech recognition [5] but also for music modelling [6]. Such pre-processing includes energy-like features (actually the zero'th cepstral coefficient) along with cepstra representing the general spectral shape.

Audio is down-sampled (if necessary) to 32 kHz and pre-emphasised to flatten the spectral shape. A 512 point DFT is performed in steps of 10 ms and squared to give the power spectrum. The resulting 257 unique bins are then decimated using a filter-bank of 23 overlapping triangular filters equally spaced on the mel-scale. The mel scale corresponds roughly to the response of the human ear. A logarithm and DFT then yield the mel-cepstrum, which is truncated, retaining the lower 13 dimensions. This truncation retains spectral shape and discards excitation frequency. Next, Cepstral Mean Normalisation (CMN) is performed by subtracting from each cepstral vector the mean of the vectors of the preceding (approximately) half second. This has the effect of removing convolutional channel effects. Finally, the 13 normalised cepstral coefficients are then augmented by first and second order derivatives, corresponding to their velocity and acceleration. This gives $k=39$ dimensional vectors.

## 3. TEMPORAL ALIGNMENT

We assume that test samples are relatively short, thus we can ignore the clock skew problem between test and reference (i.e., there is almost zero skew due to unsynchronised clocking of different devices). Presumably in some cases for long recordings the two could become misaligned, in which case additional techniques such as dynamic time warping [7] should be taken into account during the matching process. We consider two operating modes, one un-normalised and the other normalised.

### 3.1 Euclidean Mode

The temporal alignment in un-normalised Euclidean mode is performed by searching for a best distance in *n*-dimensional Euclidean space between test time-frequency matrix (corresponding to a test clip) and master time-frequency matrix with 10 ms step:

$$t_i^* = 10 \cdot \arg\min_{M_p^{(i)} \in M} (d(M_p^{(i)}, V_i))$$

$$d(M_p^{(i)}, V_i) = \sum_{q=1}^{N_i} \| m_{p+q} - v_{i,q} \| \cdot$$

In the above equation, $i$ is the index of the corresponding test clip ($1 \le i \le N$, where $N$ is the number of test clips), $t_i^*$ is the relative position in ms of $i$'th test matrix and $d$ is a Euclidean metric. $V_i$ is the $i$'th test matrix, $v_{i,q}$ is the $k$-dimensional vector of the $i$'th test matrix, which corresponds to a frame represented by k pre-processed coefficients. $N_i$ is the number of frames inside the matrix $V_i$. $M$ is the master matrix, $M_p^{(i)}$ is the $n$-dimensional sub-matrix of the master matrix, shifted from the beginning by $10p$ ms and $m_p$ is the $k$-dimensional vector of the master matrix shifted from the beginning by $10p$ ms and corresponding to a frame represented by $k$ pre-processed coefficients.

Dimension $n$ of Euclidean space is defined as the size of the test matrix $V_i$ and equals $kN_i$. It can be calculated as the length of the test clip in ms divided by 10 and multiplied by the number, $k$, of coefficients per frame.

### 3.2 Normalised Euclidean Mode

The normalised Euclidean approach has been formulated to minimise the variance of the distance space. The temporal alignment in this mode is performed by searching for a best distance in $n$-dimensional Euclidean space between normalised test time-frequency matrix (corresponding to test clip) and normalised master time-frequency matrix with 10 ms step:

$$t_i^* = 10 \cdot \arg\min_{M_p^{(i)} \in M} (d(M_p^{(i)}, V_i))$$

$$d(M_p^{(i)}, V_i) = \sum_{q=1}^{N_i} \| \alpha_{p+q} m_{p+q} - \beta_{i,q} v_{i,q} \|$$

$$\alpha_{p+q} = \begin{cases} 1, if \ \| m_{p+q} \| \le 1 \\ \dfrac{1}{\| m_{p+q} \|}, if \ \| m_{p+q} \| > 1 \end{cases}$$

$$\beta_{i,q} = \begin{cases} 1, if \ \| v_{i,q} \| \le 1 \\ \dfrac{1}{\| v_{i,q} \|}, if \ \| v_{i,q} \| > 1 \end{cases}$$

In the above formulation, the parameters are as before, except $d$ is defined by the formula above, $\alpha_{p+q}$ and $\beta_{i,q}$ are normalization coefficients for corresponding frames $p+q$ and $q$ of the matrices $M$ and $V_i$.

The elements $\alpha_{p+q}$ and $\beta_{i,q}$ are upper thresholded at 1 to decrease the impact of quiet frames.

As before, dimension $n$ of Euclidean space is defined as the size of the test matrix $V_i$ and equals $kN_i$.

## 4. EXPERIMENTAL RESULTS

Experimental results were achieved on a real life dataset of 100 recordings:

- 1 master track 51 min length recorded by the reference Canon camera (10.8GB, video: DVSD 720x576 25.00fps 28799Kbps, audio: PCM 32000Hz stereo 1024Kbps),
- 17 clips 12-130 seconds length recorded by Nokia mobile phone (0.3 GB, video: MPEG4 Video 640x480, audio: AAC 48000Hz mono 768Kbps),
- 28 clips 4-133 seconds length recorded by Canon camera (0.8GB, video: MPEG2 Video 720x576 25.00fps 9600Kbps, audio: Dolby AC3 48000Hz stereo 256Kbps),
- 15 clips 16-695 seconds length recorded by Sony camera (8.4GB, video: DVSD 720x576 25.00fps 28800Kbps, audio: PCM 48000Hz stereo 1536Kbps),
- 39 clips 1-250 seconds length recorded by Sanyo camcorder (1.2GB, video: MPEG4 Video H.264 1280x720 29.97fps, audio: AAC 48000Hz stereo 1536Kbps).

The master track content consists of a high school rehearsal with multiple events/replays one after the other. All corresponding audio tracks were extracted and converted to 32kHz mono PCM files with VirtualDub software [8].

Experiments were conducted on a closed set (i.e. we did not consider a rejection mechanism for test segments that did not correspond to the master track).

Processing time (on a Pentium 4 2.4GHz) for the algorithm without multi-core optimisation was 26 seconds for automatic temporal alignment of a 12 seconds test recording over the 51 min master track using 5 cepstra, and 60 seconds for the same segments using 12 cepstra. It is directly proportional to the length of the test segment, to the length of the master track and to the feature vector dimensionality.

To avoid possible inaccuracy associated with manual annotation the performance was calculated as the number of correctly (within ±5 frames) aligned clips divided by the total number of test clips.

## 4.1 Euclidean mode

In figure 1 we illustrate how the dimensionality of the feature vector influences total performance.

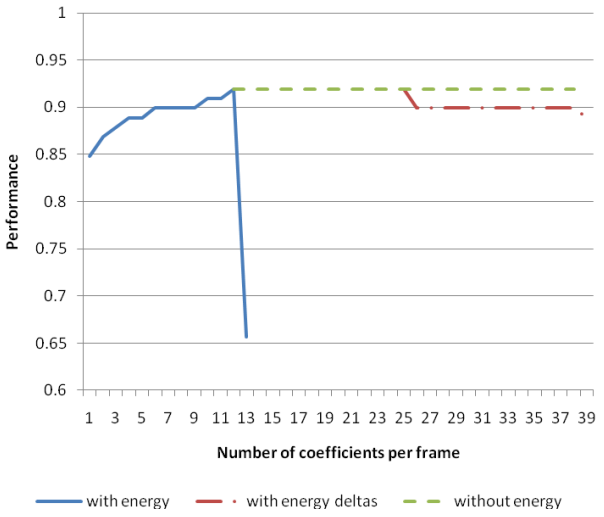


**Figure 1. Euclidean mode performance.**

It is clearly visible that the performance improves with increasing cepstral analysis order. However, it dramatically drops when the energy is considered (solid line). We suppose this drop is due to the increased variance of the search distance space. Delta features do not add any value (dashed line), we suppose due to the fact that deltas can be easily reconstructed from cepstra over time. Further, they are used in ASR as a continuity constraint, which is not necessary in this application. The use of energy deltas results in a small drop in performance (long dashed dotted line), again we suppose due to the fact that they slightly increase the variance of the search distance space. The best result is 92%.

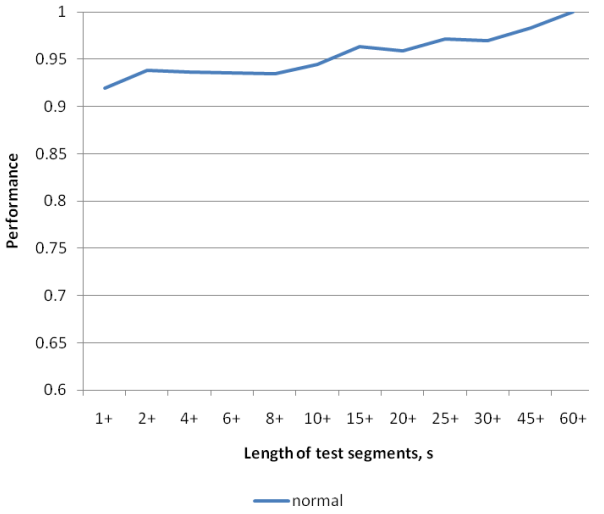


**Figure 2. Euclidean mode performance with 12 cepstra versus test segment length.**

Nevertheless, there is also a strong dependency on the length of test recordings. In figure 2 we illustrate how the length of the test segments impacts on the total performance. The performance grows and, for recordings longer than 60 s, 100% performance is achievable on the described dataset.

## 4.2 Normalised Euclidean mode

In figure 3 we illustrate how the dimensionality of the feature vector influences total performance in normalised Euclidean mode.

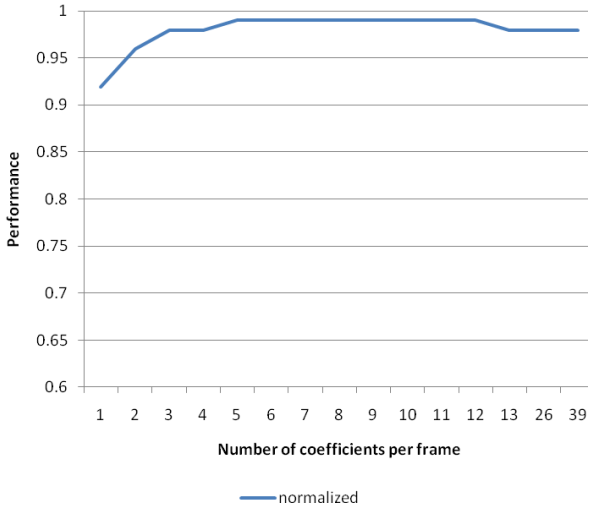


**Figure 3. Normalised Euclidean mode performance.**

The results become very stable for more than 5 cepstra, and there is small drop in performance when energy is considered. The best result is 99%.

For 5-12 cepstra, incorrect detection was observed only in cases when recordings of only 1 s in length were considered. When such recordings were omitted the performance increased to 100% on the described dataset.

More discussion concerning improved performance of this approach is presented in the following section.

## 4.3 Impacts of normalization

To better understand the achieved results we examine the distance distribution in search space (the dimension of the search space in our case is approximately 300,000 elements – equal to the length of the master track in steps of 10 ms) in the case of Euclidean mode (figure 4) and in the case of normalised Euclidean mode (figure 5). 12 cepstra are used for both cases.

The valley in the magnified area on each figure corresponds to the correct result (for the same test clip). Nevertheless, this valley is not always caught by the automatic temporal alignment algorithm in Euclidean mode due to the high variance of the search distance space. Due to the reduced variance of the search distance space in the normalised Euclidean mode the performance increases.

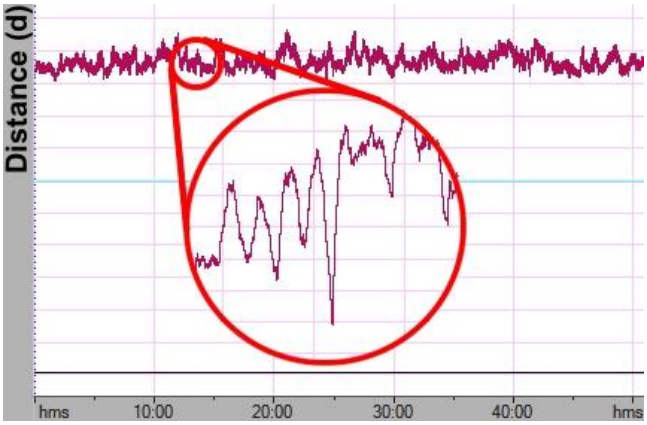


**Figure 4. Euclidean mode distance space.**


**Figure 5. Normalised Euclidean mode distance space.**

It is worth mentioning that the variance of the search space is directly proportional to the length of the test recordings. This is why, on recordings 60+ s in length, we observe very good results even in Euclidean mode.

## 5. CONCLUSION

We have shown that multiple AV signals can be aligned to an acceptable accuracy using audio features typical of ASR applications. Surprisingly, we find that the energy of the signal is not good for alignment, but that good alignment can be inferred from a small number of normalised cepstra. We have shown that results can be improved using a simple normalisation.

## 6. ACKNOWLEDGMENTS

This work was supported by the European Union FP7-ICT project TA2. We are grateful to British Telecom for provision of the real life dataset.

## 7. REFERENCES

[1] Verrier, Jean-Marc, 1999. Audio Boards and Video Synchronisation. AES UK 14th Conference: Audio - The Second Century.

[2] Dannenberg, R.B. & Hu, N., 2003. Polyphonic Audio Matching for Score Following and Intelligent Audio Editors. Proc. 2003 ICMC, ICMA.

[3] Birmingham, W.P., et al., 2001. MUSART: Music Retrieval via Aural Queries. Proc. 2nd International Symposium on Music Information Retrieval (ISMIR): 73-81.

[4] Mermelstein, P., 1976. Distance measures for speech recognition, psychological and instrumental. In Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed., pp. 374–388. Academic, New York.

[5] Darren Moore, John Dines, Mathew Magimai.-Doss, Jithendra Vepa, Octavian Cheng and Thomas Hain, 2006. Juicer: A Weighted Finite-State Transducer speech decoder. In 3rd Joint Workshop on Multimodal Interaction and Related Machine LEarning Algorithms MLMI'06.

[6] Logan, B., 2000. Mel Frequency Cepstral Coefficients for Music Modeling. Proc. 1st ISMIR.

[7] Ning Hu, Roger B. Dannenberg and George Tzanetakis, 2003. Polyphonic Audio Matching and Alignment for Music Retrieval. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 19-22, 2003.

[8] Video capture/processing utility VirtualDub. http://www.virtualdub.org/